METRUM RESEARCH GROUP

Bayesian analysis in NONMEM

Tim Waterhouse
Metrum Research Group

30 March 2023

Acknowledgments

These slides borrow heavily from the following:

- Introduction to Bayesian pharmacometric data analysis with NONMEM (ACoP 2019 workshop by Bill Gillespie and Curtis Johnston)
- "Tutorial: Bayesian Estimation in NONMEM" by Curtis K. Johnston, Matthew Wiens, John Mondick, Jonathan French, Bill Gillespie (and myself) (manuscript submitted for publication)

Bayesian analysis in NONMEM

- Why Bayesian?
- Introduction to Bayesian statistical principles and methods
- Computation for Bayesian modeling
- Overview of NONMEM® implementations
- Prior distributions
- Model evaluation and comparison
- Assessing convergence and choosing numbers of burn-in and post-burn-in samples
- Getting your hands on posterior samples for individual parameters and predictions
- When stuff goes wrong
- Practical strategies for selecting Bayesian estimation methods for specific types of problems
- Example Bayesian analysis

Why Bayesian?

- Pharmacometricians are often called on to leverage prior knowledge in order to interpret new data and facilitate decision-making in drug development.
  - Qualitative prior knowledge is captured in the mathematical form of a model, i.e., the **likelihood function**.
  - Quantitative prior knowledge may be captured in the form of probability distributions of model parameter values, i.e., **prior distributions**.
- Add **data** and you have all the ingredients of Bayesian data analysis.
- With Bayes Rule and suitable computation tools those components are combined to yield **posterior distributions** of model parameters and predictions.
- Those distributions permit probabilistic inferences directly relevant to decision-making.

Introduction to Bayesian statistical principles and methods

- Bayesian principles and methods provide a coherent framework for:
  - Quantifying uncertainty,
  - Making inferences in the presence of that uncertainty.
- It is also the basis for formal approaches to incremental model building, parameter estimation and other statistical inference as knowledge and data are accumulated.

The two core notions that distinguish Bayesian analysis are:

- Unknown quantities are viewed as random variables, i.e., they are described in terms of probability distributions.
- Bayes rule provides a formal mechanism for combining prior knowledge and new data.

Bayesian inference

Bayes Rule

Bayes Rule is the basis for inference about model parameters (θ) given data (y) and prior knowledge about model parameters ($p(\theta)$):

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)\, p(y|\theta)}{p(y)} = \frac{p(\theta)\, p(y|\theta)}{\int p(\theta)\, p(y|\theta)\, d\theta} \\ &\propto p(\theta)\, p(y|\theta) \end{aligned}$$

The p's are probabilities or probability densities of the specified random variables.

Bayesian modeling/inference process

1. Assess prior distribution $p(\theta)$
   - θ viewed as random variables
   - Subjective
   - Ideally base on all available evidence/knowledge (or belief)
   - Or deliberately select a non-informative prior (e.g., reference, vague or improper prior)
2. Construct a model for the data $p(y|\theta)$, also known as the likelihood function when viewed as a function of θ.
3. Calculate posterior distribution $p(\theta|y)$.
   - Use for inferences regarding parameter values
4. Calculate posterior predictive distribution $p(y_{\text{new}}|y)$.
   - Use for inferences regarding future observations

$$p(y_{new}|y) = \int p(y_{new}|\theta)\, p(\theta|y)\, d\theta$$

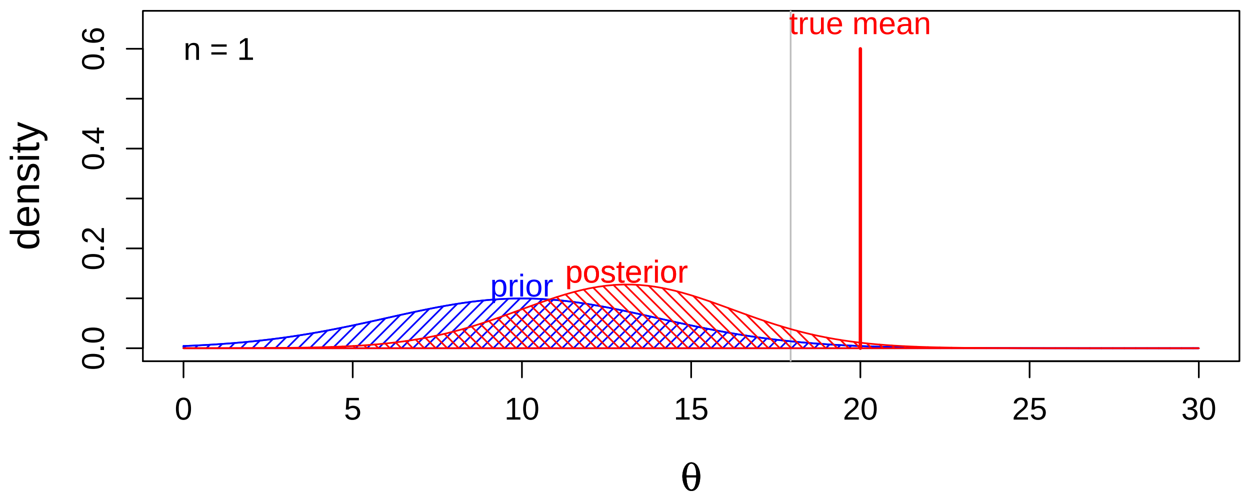
A simple one parameter example

Estimating the mean of a normal distribution with a known variance where the prior is also normal:

$$\begin{aligned}
\theta &\sim N(\mu_0, \tau_0) \qquad\qquad y|\theta \sim N(\theta, \sigma) \\
p(\theta|y) &\propto p(\theta)\, p(y|\theta) = p(\theta) \prod_{i=1}^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\right) \\
&\Downarrow \\
\theta|y &\sim N(\mu_n, \tau_n) \\
\mu_n &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}
\end{aligned}$$

A simple one parameter example

- Posterior mean is a weighted average of prior mean (μ_0) and the sample mean ($\overline{y}$).
- Posterior precision $\left(\frac{1}{\tau_n^2}\right)$ is a sum of the prior precision $\left(\frac{1}{\tau_0^2}\right)$ and the data precision $\left(\frac{n}{\sigma^2}\right)$.



Computation for Bayesian modeling

- Full Bayesian analysis requires:
  - Characterization of the joint posterior distribution of model parameters and of predicted outcomes.
  - Integration of that joint posterior distribution to calculate quantities required for statistical inferences.
- For most realistic problems, those are very computationally demanding tasks.
- Increases in computation speed and development of new algorithms over the last 25–30 years have finally made full Bayesian analysis a feasible option for routine data analysis.

Full Bayesian analysis

- Full Bayesian analysis refers to characterization of the joint posterior distribution, not just its mode.
- Most Bayesian inference is based on univariate marginal posterior distributions of individual parameters or functions of those parameters.
- That requires integration of the joint posterior distribution, usually over several dimensions.

Full Bayesian analysis

- If you're lucky those integrals have known analytic solutions, but that is rarely true for PK/PD modeling applications.
- For integrals in fewer dimensions, a numerical quadrature method might be practical.
- Now imagine the computational requirements for hierarchical models, e.g., population PK models, with individual-specific parameters in the hundreds!!

Posterior simulation

- What if you could simulate samples of θ from the joint posterior distribution?
- Then you could estimate $E\left(f\left(\theta\right)|y\right)$ by the arithmetic mean:

$$E\left(f\left(\theta\right)|y\right) \approx \frac{1}{n}\sum_{i=1}^{n} f\left(\theta_i\right)$$

- More generally, you could characterize the properties of any marginal posterior distribution of a model parameter or function of model parameters, e.g., moments, quantiles, ...
- But how do you simulate samples from a high dimensional joint posterior distribution?
- Markov chain Monte Carlo (MCMC) simulation via NONMEM is one approach we will explore today.

Markov Chain Monte Carlo (MCMC) simulation

- Involves random draws from approximate distributions and then correcting those draws to better approximate the joint posterior.
- The samples are drawn sequentially so that each draw depends on the previous one, thus forming a Markov chain.
- Eventually the Markov chain converges (in distribution) to a stationary distribution that is the joint posterior distribution.
- Algorithms for MCMC include:
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Hamiltonian Monte Carlo (HMC) simulation
- MCMC samples are serially correlated:
  - Inferences based on MCMC require more samples than would be required for independent samples
- Practical consequences:
  - Use only samples drawn after convergence is achieved, i.e., discard samples from a “warmup” phase.
  - Draw more samples than you would for independent random draws.

HMC performance

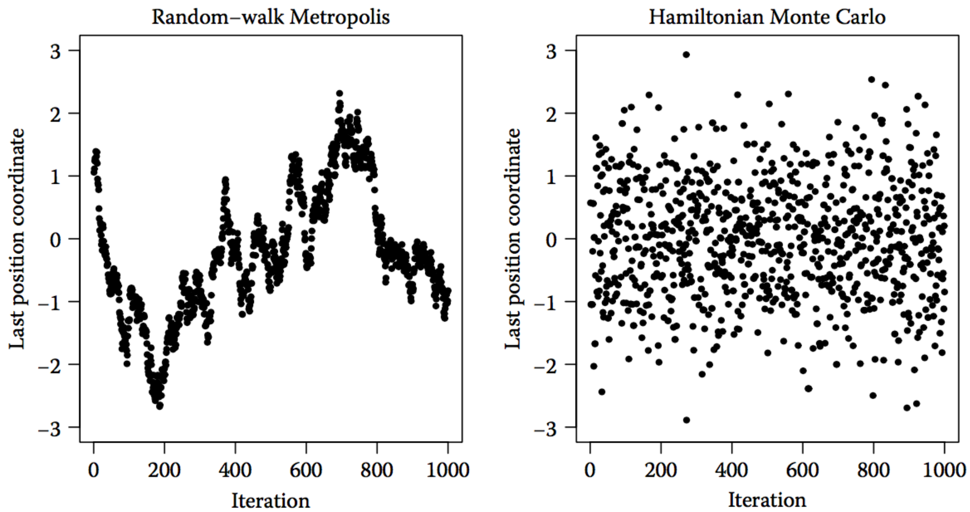


FIGURE 5.6
Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.

from RM Neal. MCMC Using Hamiltonian Dynamics (2011) [1]

HMC performance

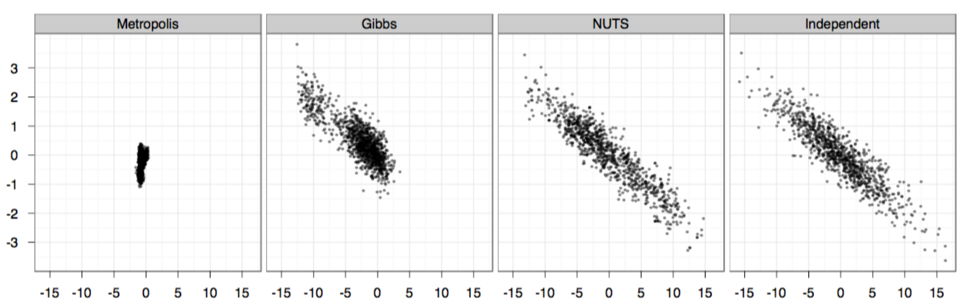


Figure 7: *Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000 samples (thinned to 1,000 samples for display) generated by Gibbs sampling (second from left), and 1,000 samples generated by NUTS (second from right). Only the first two dimensions are shown here.*

from MD Hoffman and A Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo (2014) [2]

Also see: `http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/`

Overview of NONMEM implementations

- MAP estimation
  - Using prior distributions with any optimization method
- MCMC
  - METHOD = BAYES: Metropolis-Hastings within Gibbs sampling
  - METHOD = NUTS: No U-turn sampler (HMC with automatic optimization of sampling parameters)
  - Remember:
    - MH (Metropolis-Hastings) is “meh”
    - Gibbs is “good”
    - HMC (NUTS) is “how maestros compute”

Brief discussion on prior distributions

- Think of prior distributions as part of the model.
- Priors should be chosen and subjected to scrutiny much like other model components.
- Model checking should ideally include sensitivity analysis of the priors.
- Choice of priors is most critical with sparse or limited data.

See https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

What does it mean to be informative, uninformative or weakly informative?

Not well defined, but here's an attempt at some loose definitions:

- Weakly informative prior: A prior that rules out unreasonable parameter values but is not so strong as to rule out values that might make sense
- Informative prior: A prior that purposely represents information intended to influence the posterior distribution
  - To capture prior knowledge
  - To challenge the analysis with competing points of view, e.g., use of pessimistic or optimistic priors.
- Uninformative prior: Ostensibly a prior that represents no information and therefore "let's the data tell the story."
  - E.g., a constant over the entire real line—an improper prior

Beware: That "uninformative" prior might not be!

- Suppose you use an improper prior for a standard deviation—a constant over the positive real line.
- That means all positive values are equally likely. Sounds like a reasonable definition of uninformative doesn't it?
- But that means that the prior assigns infinitely more probability to the set of values greater than any fixed value you care to choose.
- This will tend to bias the posterior to high values.
- Bottom line: A uniform distribution does not automatically confer uninformativeness.

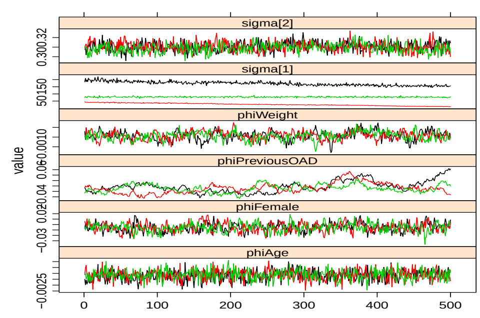
Assessing convergence & adequacy of sample sizes

- Early samples may be unrepresentative of the target distribution
- MCMC samples within a chain are autocorrelated
  - Inferences based on MCMC samples are less precise than those from the same number of independent samples
  - Autocorrelation also influences the rate of convergence

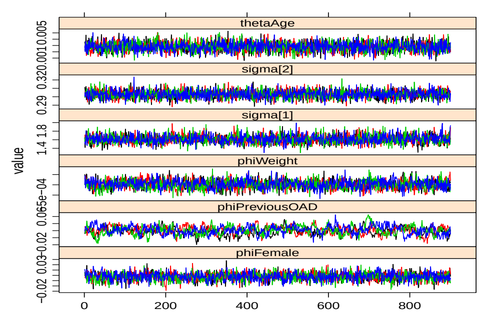
Assessing convergence & adequacy of sample sizes

- Use a warmup phase, i.e., discard early iterations
- Monitor convergence via multiple chains with different starting points
  - Look for chains to converge to a common distribution
  - You want chain history plots to look more like straight horizontal “fuzzy caterpillars” than “wiggly snakes”

Poor convergence & mixing

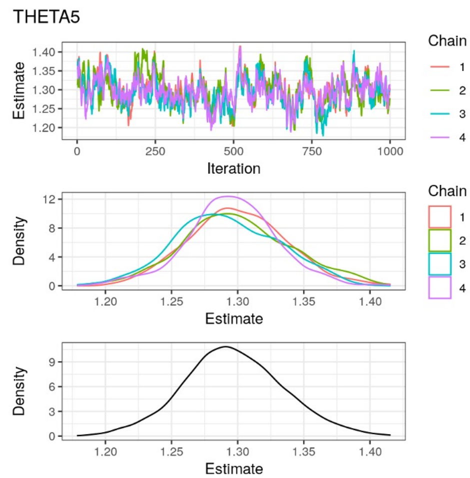


Good convergence & mixing



Assessing convergence & adequacy of sample sizes

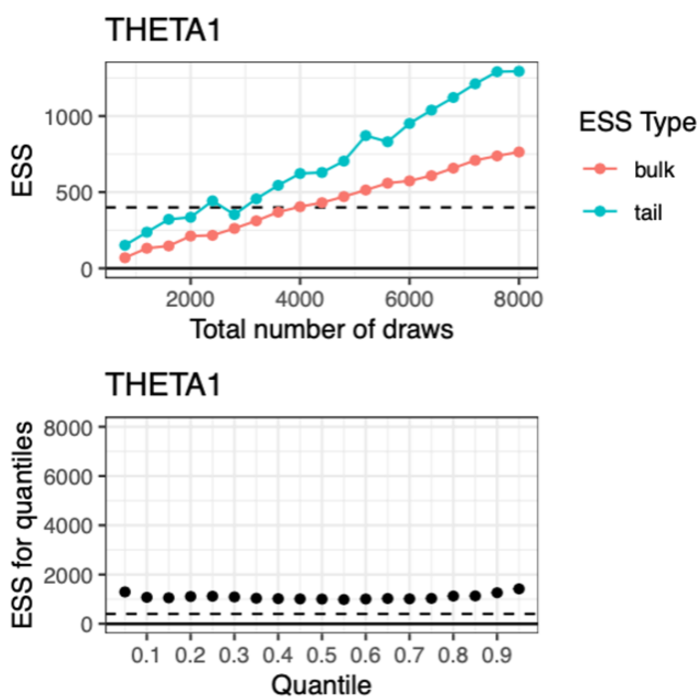
- Gelman-Rubin diagnostics (Rhat or $\hat{R}$)
  - Essentially ratios of total variance to within chain variance.
  - Should approach 1 for all parameters of interest on convergence
- Effective sample size (ESS)
  - Measure of sampling efficiency
  - Bulk (location of distribution)
  - Tail (5th and 95th percentiles of distribution)
  - Desire ESS $> \approx$400



Bulk ESS = 188
Tail ESS = 354
Rhat = 1.01

Assessing convergence & adequacy of sample sizes

- ESS vs draw
  - Will longer chains solve convergence issues?
- ESS vs quantile
  - Ensure convergence across all quantities of interest



IMHO

- Do NOT trust automatic convergence detection to determine number of burn-in samples.
- In particular, do NOT use NONMEM's termination tests (CTYPE option) for terminating burn in.

When stuff goes wrong

- Diagnosing and remedying sampling problems encountered with MCMC
- Reparameterization, e.g., centered vs non-centered parameterizations for hierarchical models
- Prior distributions as part of the solution

Improving computational efficiency and sampling performance

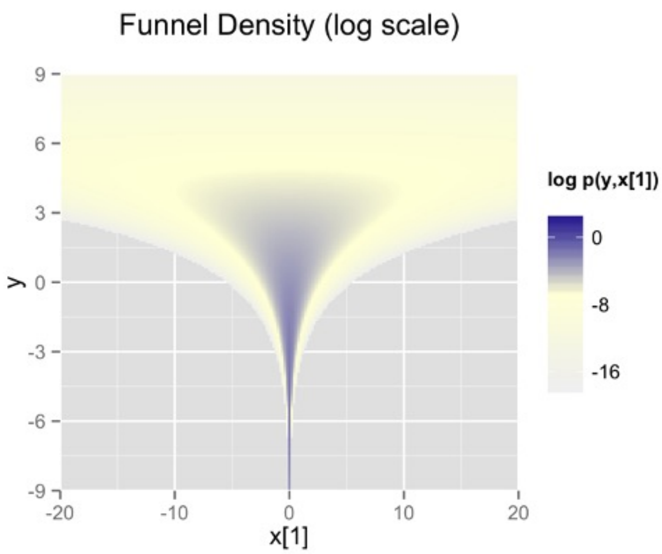
The main strategies are:

- Reparameterization
- Adjusting MCMC tuning parameters
- Weakly informative priors to regularize fitting of hierarchical models

Reparameterization

- Model parameterization can markedly affect the geometry of the joint posterior distribution.
- Posterior distributions with extreme curvature or sharp transitions lead to sampling inefficiency or outright failure.
- A well-selected reparameterization can often “smooth out” the geometry, thereby improving sampling performance.
- For example switch from centered to non-centered parameterization (or vice versa) of random effects in hierarchical models.
- Other reparameterizations that reduce posterior curvature and correlation, e.g., truncated Emax parameterization.
- Use MU referencing when possible with BAYES and NUTS.

Devil's funnel



$$y \sim N(0, 3)$$
$$x \sim N(0, \exp(y/2))$$

Centered (CP) and non-centered (NCP) parameterization

Consider an individual parameter $\phi_i \sim N(\mu, \omega)$.

- CP refers to specifying ϕ_i directly using that distribution centered at μ.
- NCP refers to specifying ϕ_i indirectly using a standard normal, $\eta_{\text{std},i} \sim N(0, 1)$, and then calculating $\phi_i = \mu + \omega\eta_{\text{std},i}$
- The above is generalized to multivariate normal case using a Cholesky decomposition to generate a multivariate normal vector from standard normal ηs.

Centered (CP) and non-centered (NCP) parameterization in NONMEM

- For NUTS,
  - AUTO = 0 or 1 for CP
  - AUTO = 2 for NCP
  - AUTO = 3 for something in between. Uses $N(0, \omega)$ instead of $N(0, 1)$
- To my knowledge there are no equivalent implementations for BAYES.

Reparameterize to reduce correlation among parameters

PII-108

"TRUNCATED SIGMOID E_{max} MODELS": A REPARAMETERIZATION OF THE SIGMOID E_{max} MODEL FOR USE WITH TRUNCATED PK/PD DATA. WJ Bachman PhD and WR Gillespie PhD, GloboMax LLC, Hanover, MD.

The parameters of the sigmoid E_{max} model are poorly estimated when the range of PK/PD data available is limited to $<0.95E_{max}$ [Dutta et al. J Pharm Sci 85:232 (1996)]. The following reparameterized form of the sigmoid E_{max} model has improved parameter estimation properties:

$$E = E_0 + \frac{(\beta^\gamma + 1)(E^* - E_0)C^\gamma}{C^{*\gamma} + \beta^\gamma C^\gamma}$$

where E is the effect measure and C is a measure of drug exposure (e.g., concentration or dose). The parameter E^* is the estimated effect resulting from C^*, γ is the usual "sigmoidicity" parameter, and E_0 is the baseline effect. β is a measure of the degree to which the function deviates from linearity in C^γ. One approach to applying this parameterization is to fix C^* (or E^*) at a value and estimate the remaining parameters E_0, E^* (or C^*), β, and γ by nonlinear regression. The properties of this approach are evaluated by application to simulated PK/PD data that is truncated at various fractions of E_{max}. When C^* (or E^*) is chosen within the range of the observed data, then the parameters E^* (or C^*) and β are more precisely and accurately estimated than EC_{50} and E_{max} of the standard parameterization.

CP&T 63:199 (1998)

- With BAYES method high correlation among parameters often causes high autocorrelation in the MCMC samples.
- This often happens with asymptotic functions like the Emax function.
- NUTS is more robust to such correlation.

Prior distributions as part of the solution

IIV variances are difficult to estimate, particularly with data from a small number of individuals. What to do?

- Reduce the number of random effects until you find a set you can estimate with high precision, or
- Use a weakly informative prior for Ω that is consistent with our knowledge of IIV and excludes clearly implausible values.

I argue that for most PMX models the latter is more consistent with our knowledge and should be the preferred approach.

Other considerations for NONMEM

- Control stream
  - Priors
  - Initial estimates
- Diagnostics
  - MCMC diagnostics
  - Model diagnostics

NONMEM control stream

- MU reference when possible
  - Allow Gibbs sampling (vs MH) for `METHOD=BAYES`
  - Analytic derivatives for `METHOD=NUTS`
- Prefer unbounded THETAs
  - Log or logit transform where possible
- Specify as many priors as possible

NONMEM control stream: priors for THETAs

- Normal distribution
  - Mean `$THETAP`
  - Variance `$THETAPV`
    - Shorthand:
      `$THETAPV BLOCK(5) FIXED VALUES(10,0)`
- t-distribution (`METHOD=NUTS`)
  - Set degrees of freedom in `$EST TTDF` or `$TTDF`

NONMEM control stream: priors for OMEGAs

- Inverse Wishart distribution
  - Mode `$OMEGAP`
  - Degrees of freedom `$OMEGAPD`
- Additional options for `METHOD=NUTS`:
  - Lognormal or half-t-distribution for SDs (`$EST OVARF`)
  - Lewandowski-Kurowicka-Joe (LKJ) distribution for correlation matrix (`$EST OLKJDF`)

NONMEM control stream: priors for SIGMAs

- Inverse Wishart distribution
  - Mode `$SIGMAP`
  - Degrees of freedom `$SIGMAPD`
- Additional options for `METHOD=NUTS`:
  - Lognormal or half-t-distribution for SDs (`$EST SVARF`)
  - Lewandowski-Kurowicka-Joe (LKJ) distribution for correlation matrix (`$EST SLKJDF`)

NONMEM parameterization of the inverse Wishart distribution

NONMEM implementation of inverse Wishart prior for Ω (or Σ):

$$\begin{aligned}
\Omega &\sim W^{-1}\left(\nu\Omega_{\text{prior}}, \nu\right) \\
W^{-1} &= \text{Standard parameterization of the inverse Wishart} \\
\nu &= \text{degrees of freedom} \\
E\left(\Omega\right) &= \frac{\nu\Omega_{\text{prior}}}{\nu - n - 1} \\
\text{Var}\left(\Omega_{ii}\right) &= \frac{E\left(\Omega_{ii}\right)\left(\nu - n - 1\right)}{\nu} \\
n &= \dim\left(\Omega\right)
\end{aligned}$$

Inverse Wishart distribution

Guidance for setting $Omega_{\text{prior}}$ and ν:

$$\begin{aligned}
\nu_i &= \frac{2E\left(\Omega_{ii}\right)^2}{\text{Var}\left(\Omega_{ii}\right)} + n + 3 \\
\Omega_{\text{prior},ii} &= \frac{E\left(\Omega_{ii}\right)\left(\nu_i - n - 1\right)}{\nu_i}
\end{aligned}$$

- Set diagonal elements of Ω_{prior} to the calculated values of $\Omega_{\text{prior},ii}$
- Set $\nu = \min\left(\nu_i\right)$.

Estimation options: initial estimates

- Multiple (e.g., 4) chains using `METHOD=CHAIN`.
  - Generate 4 sets of initial estimates with `METHOD=CHAIN NSAMPLE=4 FILE=1000.chn`
  - Use `CTYPE` option to sample initial THETAs from
    - uniform (% above and below `$THETA`), or
    - bounds in `$THETA` (not recommended!), or
    - normal distribution defined by `$THETAP` and `$THETAPV`
  - OMEGA and SIGMA initial estimates from inverse Wishart distributions
    - Degrees of freedom from `DF` and `DFS`

Bayesian diagnostics in NONMEM

- Diagnostics should consider full posterior (across all chains)
- NONMEM generates summaries (means, standard errors, shrinkages, etc.) **within each chain**
- Further post-processing is required to summarize and diagnose models **across all chains**

NONMEM output needs some tweaking

- NONMEM will output means of parameter estimates
  - Probably OK for THETAs, can introduce bias for variance terms
- $TABLE outputs
  - ETA values not derived across posterior distribution, but post hoc estimates using mean of THETAs/OMEGAs
    - May result in spurious correlations

Better to derive estimates/diagnostics using the full posterior: PREDs and IPREDs

- Simulate S replicates:
  - $y_{ijs}^{\text{sim,PRED}}$
    - include all variability, sample from posterior at each replicate
  - $y_{ijs}^{\text{sim,IPRED}}$
    - include within-subject variability, posterior samples of population parameters and ETAs
  - Calculate:
    - $\text{PRED}_{ij} = \frac{1}{S}\sum_{s=1}^{S} y_{ijs}^{\text{sim,PRED}}$
    - $\text{IPRED}_{ij} = \frac{1}{S}\sum_{s=1}^{S} y_{ijs}^{\text{sim,IPRED}}$

Better to derive estimates/diagnostics using the full posterior: Shrinkage

- Shrinkage = $1 - \frac{\mathrm{SD}_{k=1}^{K}(\overline{\eta_k})}{\sqrt{\overline{\Omega}}}$
  - $\overline{\eta_k}$ is mean of ETA posterior samples for subject k
  - $\mathrm{SD}_{k=1}^{K}$ is standard deviation across K subjects
  - $\overline{\Omega}$ is mean of OMEGA estimates across posterior samples

Better to derive estimates/diagnostics using the full posterior: NPDE

- Can be calculated with npde R package
- Reuse output from PRED simulations: $y_{ijs}^{\mathrm{sim,PRED}} \Rightarrow Y_i^{\mathrm{sim}(k)}$

$$\mathbf{Y}_i^{\mathrm{sim}(k)*} = \mathrm{var}(\mathbf{Y}_i)^{-1/2}(\mathbf{Y}_i^{\mathrm{sim}(k)} - E(\mathbf{Y}_i))$$

$$\mathbf{Y}_i^{*} = \mathrm{var}(\mathbf{Y}_i)^{-1/2}(\mathbf{Y}_i - E(\mathbf{Y}_i))$$

$$\mathrm{pde}_{ij} = F_{ij}^{*}(y_{ij}^{*}) \approx \frac{1}{K}\sum_{k=1}^{K}\delta_{ijk}^{*}$$

where $\delta_{ijk}^{*} = 1$ if $y_{ij}^{\mathrm{sim}(k)*} < y_{ij}^{*}$ and 0 otherwise.

[5]

Model selection criteria: what not to use

- Traditional objective function comparison not appropriate
- Alternatives: AIC, DIC, WAIC, cross-validation
  - AIC: not suitable for strong informative priors
  - DIC: unreliable for non-Gaussian posteriors
  - WAIC: not robust with weak priors or influential observations
  - Cross-validation: too computationally demanding

Model selection: Use PSIS-LOO

- PSIS: Pareto smoothed importance sampling
- LOO: leave-one-out cross-validation

$$\widehat{\text{elpd}}_{\text{psis-loo}} = \sum_{i=1}^{n} \log \left(\frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s} \right)$$

- Available using `loo` R package
- $p(y_i|\theta^s)$ is likelihood for a subject or observation at a given posterior sample
  - Requires post hoc calculation with posterior ETAs

Getting your hands on posterior samples for individual parameters and predictions

- With `METHOD = BAYES` or `NUTS` `NONMEM` writes MCMC sampled parameter values to a file named in the `$ESTIMATION` statement (Defaults to <model name>.ext).
- Only population parameters are written.
- Before NONMEM 7.5: Use `FORTRAN WRITE` statements (verbatim code) to write individual parameter or prediction values to files.
- NONMEM 7.5: Use `$EST ... BAYES_PHI_STORE=1` to generate <model name>.iph

When to go Bayes (and why)?

- Bayesian data analysis and inference are always applicable to PMX modeling applications.
  - I think most scientists tend to interpret probability in Bayesian terms, so it makes sense to use methods consistent with that.
  - However, BDA is usually more computationally demanding and time consuming.
- Pragmatism leads us to go Bayesian when it pays off the most.
- E.g., when you want your inferences to be informed by both prior information and new data.
- IMHO we should be using informative prior distributions more extensively, particularly during the “learn” stages of drug development.

Further reading

- Introduction to Bayesian pharmacometric data analysis with NONMEM (ACoP 2019 workshop by Bill Gillespie and Curtis Johnston)
- Bauer RJ. NONMEM Tutorial Part II: Estimation Methods and Advanced Examples. *CPT Pharmacometrics Syst Pharmacol*. 2019;8(8):538-556. doi:10.1002/psp4.12422
- Supplementary code for NONMEM Bayes tutorial paper (submitted): `https://github.com/metrumresearchgroup/NMBayesTutorial`

Example PK model

- 2-compartment PK model
- Several covariates on clearance
- GitHub repo with models/scripts (and these slides): `https://github.com/timwaterhouse/iu-nonmem-bayes-2023`

References I

[1] Radford M. Neal.
MCMC using hamiltonian dynamics.
In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman & Hall/CRC, Boca Raton, FL, 2011.

[2] Matthew D Hoffman and Andrew Gelman.
The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.
The Journal of Machine Learning Research, 15(1):1593–1623, 2014.

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin.
Bayesian Data Analysis.
CRC Press, Boca Raton, FL, third edition, 2014.

[4] Aki Vehtari, Andrew Gelman, and Jonah Gabry.
Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models.
arXiv preprint arXiv:1507.04544, 2015.

References II

[5] Emmanuelle Comets, Karl Brendel, and France Mentré.
Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R.
Comput. Methods Programs Biomed., 90(2):154–166, May 2008.